Avoiding Missing Data Biases in Phylogenomic Inference: An Empirical Study in the Landfowl (Aves: Galliformes)

Peter A. Hosner,*[1] Brant C. Faircloth,[2] Travis C. Glenn,[3] Edward L. Braun,[1] and Rebecca T. Kimball[1]

[1]Department of Biology, University of Florida

[2]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge

[3]Department of Environmental Health Science, University of Georgia

***Corresponding author:** E-mail: hosner@ufl.edu.

Associate editor: Claudia Russo

Abstract

Production of massive DNA sequence data sets is transforming phylogenetic inference, but best practices for analyzing such data sets are not well established. One uncertainty is robustness to missing data, particularly in coalescent frameworks. To understand the effects of increasing matrix size and loci at the cost of increasing missing data, we produced a 90 taxon, 2.2 megabase, 4,800 locus sequence matrix of landfowl using target capture of ultraconserved elements. We then compared phylogenies estimated with concatenated maximum likelihood, quartet-based methods executed on concatenated matrices and gene tree reconciliation methods, across five thresholds of missing data. Results of maximum likelihood and quartet analyses were similar, well resolved, and demonstrated increasing support with increasing matrix size and sparseness. Conversely, gene tree reconciliation produced unexpected relationships when we included all informative loci, with certain taxa placed toward the root compared with other approaches. Inspection of these taxa identified a prevalence of short average contigs, which potentially biased gene tree inference and caused erroneous results in gene tree reconciliation. This suggests that the more problematic missing data in gene tree–based analyses are partial sequences rather than entire missing sequences from locus alignments. Limiting gene tree reconciliation to the most informative loci solved this problem, producing well-supported topologies congruent with concatenation and quartet methods. Collectively, our analyses provide a well-resolved phylogeny of landfowl, including strong support for previously problematic relationships such as those among junglefowl (*Gallus*), and clarify the position of two enigmatic galliform genera (*Lerwa*, *Melanoperdix*) not sampled in previous molecular phylogenetic studies.

Key words: **bias, coalescent, *Coturnix*, *Gallus*, *Meleagris*, ultraconserved elements.**

Introduction

A major challenge in phylogenetics is resolving historical relationships when little time has passed between speciation events, because few DNA substitutions accrue during short time intervals (Braun and Kimball 2001; Whitfield and Lockhart 2007; Moyle et al. 2009; Wagner et al. 2013). Those few characters that do support the correct topology may then be subsequently overwritten, obscuring phylogenetic signal (Philippe et al. 2011; Patel et al. 2013). Genome-wide sequence data sets have shown great promise in resolving challenging short internal nodes in phylogenetic trees by providing millions of nucleotides and thousands of unlinked loci suitable for analyses (Faircloth et al. 2012; Lemmon EM and Lemmon AR 2013; McCormack and Faircloth 2013; Jarvis et al. 2014; Misof et al. 2014; Prum et al. 2015). However, analyses of these large, heterogeneous, genome-scale data sets are complicated— especially in light of the discordant phylogenetic signal contained within independent loci (Maddison 1997; Kubatko and Degnan 2007; Edwards 2009; Kimball et al. 2013).

A confounding problem is that evolutionary events leading to series of short internodes potentially produce an area of tree space, known as the anomaly zone, where the most common gene trees conflict with the true species tree (Degnan and Rosenberg 2006, 2009; Kubatko and Degnan 2007; Degnan 2013). Phylogenetic inference of concatenated alignments may be positively misleading in the anomaly zone (Kubatko and Degnan 2007; Roch and Steel 2014), and one common solution to this problem requires analyses in multispecies coalescent frameworks (Edwards et al. 2007; Liu and Edwards 2009). Although the theoretical advantages of coalescent approaches are clear, the strengths and weaknesses of available coalescent-based approaches when applied to empirical data remain murky (Reid et al. 2014; Liu et al. 2015; Roch and Warnow 2015; Tonini et al. 2015; Warnow 2015). Moreover, it is uncertain how often systematic error in estimates of species trees can be attributed to anomalous gene trees instead of other potential sources of error in phylogenetic inference, such as model violations or character limitation (Rannala and Yang 2008; Nabholz et al. 2011; Salichos and Rokas 2013; Betancur-R et al. 2014; Gatesy and Springer 2014; Jarvis et al. 2014).

A specific issue when analyzing phylogenomic data is deciding upon and justifying appropriate thresholds for locus inclusion and the presence of missing data (Philippe 2004; Wiens and Morrill 2011; ; McCormack et al. 2013; Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016).

Article

There are three root causes for missing data in phylogenomic alignments: 1) Stochasticity inherent in collecting data across thousands of loci, where not all loci are detected in all genomic libraries; 2) variable sequence yield among sample libraries leading to missing data across alignments; and 3) biological processes including insertions, deletions, and other chromosomal changes. Ideally, samples with lower overall yields can be reprepared and resequenced (reducing the first two types of missing data above), but in practice sample availability, expense, and project timelines often limit resequencing, especially when researchers must rely on low molecular weight historical DNA for some taxa (Mundy et al. 1997; Knapp and Hofreiter 2010). To take full advantage of the statistical power that accompanies thousands of unlinked loci and millions of nucleotides, pragmatically, some proportion of missing data must be permitted when alignments grow beyond a few focal taxa. Thus, there is a direct tradeoff between increasing locus number and/or total alignment size, which may aid phylogenetic inference, and increasing proportions of missing data, which may hinder it (Philippe 2004; Wiens and Morrill 2011; Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016).

Effects of missing data on phylogenetic studies have been a major focus of research, including work at phylogenomic scales, but most studies have concentrated on maximum likelihood (ML) and Bayesian inference of concatenated sequence alignments (Philippe 2004; Sanderson et al. 2010; Wiens and Morrill 2011; Roure et al. 2012; Streicher et al. 2016). In general, analyses of concatenated data are robust to large proportions of missing data (Burleigh et al. 2015), as long as the data matrix contains sufficient overlap among taxa to find the true tree (Sanderson et al. 2010). Drawbacks of including sites/loci with missing data in alignments include increased risk of systematic error and increased computation time without concomitant improvement in results (Rannala and Yang 2008; Kumar et al. 2012; Lemmon AM and Lemmon ER 2013; Simmons 2014). However, other issues such as taxon sampling and model choice often appear to be more important than the amount of missing data in analyses of concatenated alignments (Roure et al. 2012).

Multispecies coalescent methods may be more susceptible to the negative effects of missing data than concatenated approaches, although this idea remains largely untested, especially with empirical data (Edwards 2009; Leaché and Rannala 2011; Wiens and Morrill 2011). Coalescent models have the benefit of accommodating differing phylogenetic signals among loci due to incomplete lineage sorting (ILS), and will in theory infer correct species trees in the anomaly zone where concatenation is positively misleading (Liu and Edwards 2009; Roch and Steel 2014, but see also Sun et al. 2014; Warnow 2015). Although theoretically desirable, many multispecies coalescent approaches have severe limitations when applied to genome-scale data. Simultaneous estimation of gene trees and species trees (Edwards et al. 2007; Heled and Drummond 2010) performs well in simulations (Leaché and Rannala 2011) but that approach is computationally intractable for large numbers of taxa and thousands of loci. Gene tree reconciliation analyses are computationally feasible when applied to genome-scale data (Liu and Edwards 2009; Mirarab, Reaz, et al. 2014; Roch and Warnow 2015), but estimating reliable gene genealogies to input to these methods is difficult due to character limitation and other problems (Rosenfeld et al. 2012; Gatesy and Springer 2014; Mirarab, Bayzid, Boussau, et al. 2014; Mirarab, Bayzid, and Warnow 2014; Springer and Gatesy 2016). Thus, it is possible that gene tree reconciliation methods are more sensitive to the ill effects of missing data than concatenation. Another intriguing option are quartet-based analyses that are consistent under ILS, but bypass the problematic stage of gene tree estimation (DeGiorgio and Degnan 2010; Chifman and Kubatko 2014). These approaches have received little testing with simulated or empirical data, so their relative performance when compared with concatenated ML and other coalescent methods are unclear (DeGiorgio et al. 2014; Sun et al. 2014).

Target capture of conserved genomic regions (Faircloth et al. 2012; Lemmon et al. 2012) combined with massively parallel sequencing produce data matrices containing thousands of unlinked loci distributed across the genome suitable for phylogenetic inference. These markers can be generated efficiently, cost-effectively, and are useful across deeper evolutionary scales than restriction enzyme–based reduced-representation libraries (Rubin et al. 2012). One class of conserved genomic regions, ultraconserved elements (UCEs), has been used for phylogenetic reconstruction at a variety of scales in vertebrates, from resolving relationships among major tetrapod lineages (Crawford et al. 2012, 2015; McCormack et al. 2012, 2013; Faircloth et al. 2013; Green et al. 2014; Jarvis et al. 2014; Streicher et al. 2016) to fine-scale vertebrate phylogeography (Smith et al. 2014). UCEs feature a conserved core region with low variation (Bejerano et al. 2004) flanked on each side by more variable sites. Conserved core regions are useful as probe targets, whereas flanking regions are variable and useful for inferring historical relationships (Faircloth et al. 2012).

Target capture is efficient, but locus recovery depends on probe design, sample quality, and variables such as genome size and repeat content that affect the library preparation and/or enrichment procedures (Knapp and Hofreiter 2010; Mamanova et al. 2010; Faircloth et al. 2012). Following target capture, entire loci will be missing for some taxa due to stochastic and biological factors. Thus, if the goal of a study is to produce a complete matrix at the locus level, adding taxa will reduce the number of completely sampled loci available for inclusion. An alternative strategy is to allow inclusion of loci that lack sequence data for a subset of taxa, which increases matrix size and locus count substantially, but at the cost of increasing missing data for some taxa (Streicher et al. 2016). Here, we designate a pattern of missing data where entire loci are missing for certain taxa as "type I" missing data (fig. 1). Studies using UCEs have favored complete or mostly complete matrices at the locus level, minimizing type I missing data (Crawford et al. 2012; Faircloth et al. 2013; Smith et al. 2014; Sun et al. 2014). Recently, Streicher et al. (2016) explored the effects of including large proportions of type I missing data in UCE phylogenomic analyses. In concatenated frameworks, they found that support increased substantially with

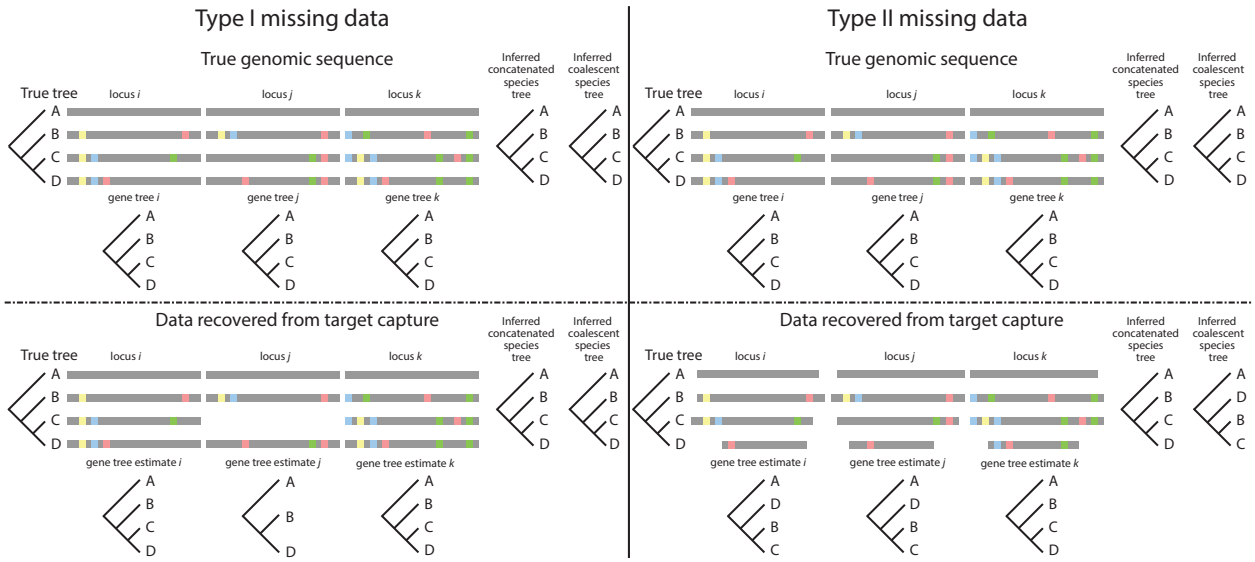

Fig. 1. A hypothetical example of how missing data within loci may bias gene tree estimation when contig lengths are shorter on average for a given taxon throughout a multilocus data set. In this example, taxon D has shorter contig lengths for loci *i*, *j*, and *k*. In two of the three gene trees, taxon D is mistakenly placed sister to B + C (because of incomplete character sampling), when rooting to outgroup taxon A. The downstream result is that taxon D is placed sister to taxon B + taxon C in gene tree reconciliation analyses, rather than sister to taxon C as in the true tree. In this example, when data are complete (true genomic sequence), or when concatenation is used, the correct tree is inferred.

increased taxon sampling and type I missing data. In coalescent frameworks, they found slightly greater support when type I missing data were lowest; however, all coalescent results suffered low bootstrap support.

Another characteristic of target capture is that the length of contigs recovered varies substantially among samples. Samples with more efficient sequence capture and greater depth of sequencing coverage typically produce longer contigs, on average, across the data set (Faircloth et al. 2013; McCormack et al. 2013). Because the majority of informative sites are found in flanking regions of UCEs (Faircloth et al. 2012), taxa having shorter average contig lengths (low N50) contain not only more missing data cells but also fewer informative sites relative to taxa having longer average contig lengths. Here, we designate this pattern of missing data where certain taxa have partial sequences for certain loci as "type II" missing data. Type II missing data could lead to errors when inferring individual gene trees for downstream analysis using gene tree reconciliation techniques (fig. 1; Simmons 2014). This potential bias caused by variation in average contig length is not limited to target capture methods or UCEs; variation in contig lengths across whole genome assemblies or other reduced representation genomic sequencing efforts can have similar effects. Type II missing data are rarely considered in large-scale phylogenies, and it is little understood how it may affect empirical phylogenetic inference of groups of interest.

The avian order Galliformes (landfowl) includes the most agriculturally important birds (the chicken [*Gallus gallus*] and turkey [*Meleagris gallopavo*], as well as Japanese quail [*Coturnix japonica*] and guineafowl [*Numida meleagris*]). The chicken is one of the premier model systems for developmental biology (Le Douarin and Dieterlen-Lièvre 2012; Hirst and Marcelle 2015) and several different galliform taxa have been used extensively in behavioral research (e.g., peafowl [*Pavo* spp.], pheasants [*Phasianus*] and relatives, and grouse [Tetraoninae]; see also Kimball et al. 2011). Other landfowl are economically important game species (e.g., ring-necked pheasant [*Phasianus*], several partridge species [*Perdix perdix, Alectoris chukar*], and northern bobwhite (*Colinus virginianus*)]. As a group, landfowl are disproportionately threatened by habitat destruction and unregulated overharvest—approximately 10% of galliform species are listed as endangered/critically endangered on the IUCN Red List (BirdLife International 2012). Yet, despite their importance in many areas of research, the galliform phylogeny is poorly resolved at many key nodes (Wang et al. 2013; Kimball and Braun 2014). For example, the identity of the sister taxon/group of *Ga. gallus*, arguably the best studied bird species in the world, still has not been resolved with confidence (Wang et al. 2013; Kimball and Braun 2014; Meiklejohn et al. 2014).

Two factors appear to limit progress toward inferring a robust galliform tree of life. First, galliforms appear to have undergone successive rapid radiations, and previous multilocus studies lacked appropriate resolution and produced conflicting results from concatenated and coalescent approaches (reviewed by Wang et al. 2013, Kimball and Braun 2014). Second, an inordinate number of galliform species are threatened by habitat destruction and overharvest (BirdLife International 2012) limiting availability of fresh tissues for key species of evolutionary interest (i.e., those not bred in captivity). Fortunately, choosing UCEs as phylogenomic markers address both these major problems in reconstructing galliform phylogeny: UCEs have demonstrated sufficient phylogenetic signal to resolve several nodes in a galliform clade that were ambiguous in previous multilocus

Table 1. Sequence Summary Statistics Comparing Concatenated UCE Matrices of Five Thresholds of Completeness.

	Alignment				
	0% Missing Taxa in Loci	<5% Missing Taxa in Loci	<25% Missing Taxa in Loci	<50 Missing Taxa in Loci	Total Evidence
Loci	140	1,740	3,361	3,919	4,817
Informative loci	140	1,739	3,329	3,868	4,638
Base pair	75,998	838,164	1,416,592	1,573,308	2,208,355
Variable sites	16,623	171,986	270,380	289,840	363,562
Informative sites	10,506	105,886	161,224	170,913	179,676
Average bootstrap %	95.0	98.8	98.6	98.8	98.8
Partitions	7	28	27	45	34
% Missing sites	10%	12%	15%	18%	39%
% Informative sites	14%	13%	12%	11%	8%
% Variable sites	22%	21%	19%	19%	17%
Average locus length	536	475	417	397	452

studies (Sun et al. 2014), and target capture shows promise in gathering thousands of loci from historical museum specimens when fresh tissues are unavailable (Knapp and Hofreiter 2010; Sun et al. 2014; McCormack et al. 2015).

The difficulties inherent in estimating galliform phylogeny (reviewed in Wang et al. 2013, Kimball and Braun 2014) also brand it a model system for understanding the consequences of methodological choices in exceptionally challenging phylogenetic inference scenarios. Here, we used UCEs to reconstruct a phylogenetic hypothesis for Galliformes, with a focus on resolving problematic nodes highlighted by previous studies. To have confidence in any resulting galliform phylogeny, it is also important to understand how both type I and type II missing data affect performance of phylogenetic methods—gene tree reconciliation frameworks in particular. We utilized model-based concatenated ML (RAxML; Stamatakis et al. 2014), two quartet methods expected to be consistent in the anomaly zone (supermatrix rooted triples, SMRT-ML; DeGiorgio and Degnan 2010; singular value decomposition scores for species quartets, SVDquartets; Chifman and Kubatko 2014), and two gene tree reconciliation approaches (Accurate Species TRee Algorithm, ASTRAL; Mirarab, Reaz, et al. 2014; Accurate Species Trees from Internode Distances, ASTRID; Vachaspati and Warnow 2015) to estimate the phylogeny of Galliformes. We apply all methodologies to five thresholds of matrix completeness (no type I missing data, 5% type I missing data, 25% type I missing data, 50% type I missing data, and total evidence), to understand how missing data influence concatenated and multispecies coalescent phylogenomic inference, particularly with respect to samples with reduced sequence yield and relatively large proportions of missing data. If results from differing methodologies and thresholds of missing data are qualitatively similar, we would conclude that methodologies are robust and missing data are of little practical consequence. However, if certain analytical techniques or missing data thresholds produce alternate strongly supported results with respect to taxa with large amounts of type I or type II missing data, they are likely biased, demonstrating that missing data are of concern.

Results

Sequence Capture Yields Data for Thousands of UCE Loci

We obtained data for 4,817 UCE loci, of which 4,638 contained at least one parsimony informative site (table 1). The number of UCE loci obtained varied from 1,035 to 4,328 per taxon, and base pairs recovered varied from 345,856 bp (16% of total aligned nucleotides) to 1,625,610 bp (74% of total aligned nucleotides) per taxon (supplementary table S1, Supplementary Material online). The large amount of data available allowed construction of data matrices up to 2.2 Mbp containing more than 180,000 informative sites. As matrix size increased in length, we observed the expected increase in number of variable and informative sites. However, the percentage of variable and informative sites decreased with increased matrix size, demonstrating a pattern of diminishing returns associated with increased missing data in larger alignments. The number of partitions identified by PartitionFinder varied from 7 to 45 depending on alignment, with larger alignments justifying greater numbers of partitions (table 1).

UCE average contig lengths recovered from sequence captures varied from 226 to 386 bp (supplementary table S1, Supplementary Material online). UCE contigs derived from historical DNA (toepads from museum specimens) were significantly shorter (unpaired *t*-test, $P < 0.0001$) in contig length (N50 311 bp) than contigs derived from fresh tissue samples (N50 372 bp). UCE enrichments using toepad source material produced similar numbers (unpaired *t*-test, $P = 0.4492$) of UCE contigs on average per enrichment (3,301 loci) to fresh material (3,654 loci).

Concatenated Maximum-Likelihood Phylogenetic Inference

ML phylogenetic inference using concatenated data sets produced strongly supported and congruent phylogenetic hypotheses for Galliformes (fig. 2). All well-supported nodes were congruent across all five alignments of varying

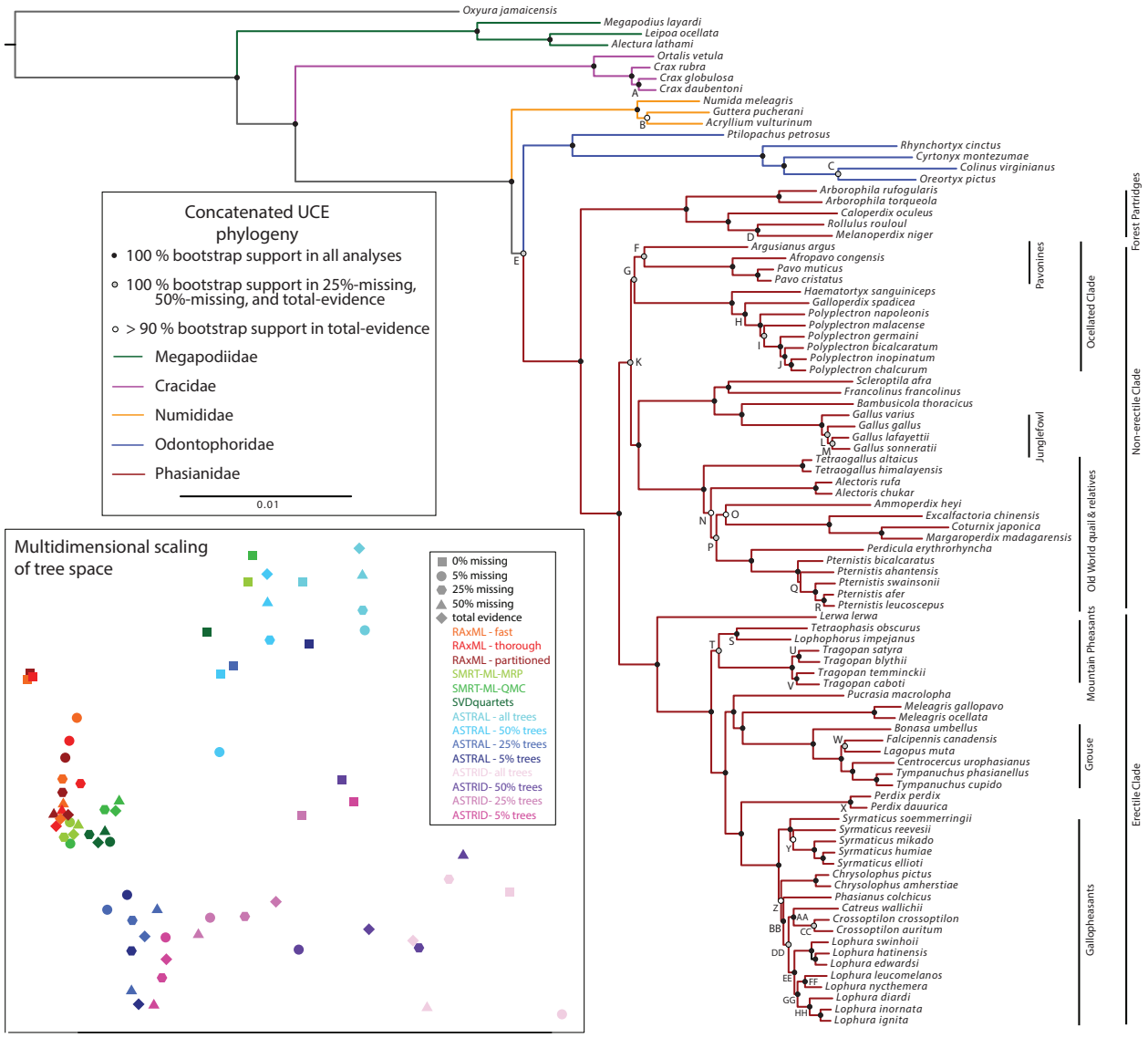

Fig. 2. Phylogeny of 90 galliform taxa inferred with ML analysis of 2,208,355 bp from 4,817 concatenated UCE loci. Inset shows multidimensional-scaled visualization of tree space, with each point representing a consensus tree produced with different inferential procedures and different thresholds of missing data (70 iterations, total). Concatenated ML, SMRT-ML, and SVDquartets trees (with the exception of those inferred from the small 0% missing data set) and ASTRAL/ASTRID trees with the most resolved gene trees converge on the same area of multidimensional tree space (lower right); RF distances between these analyses were generally <5.

completeness, with 67 of 90 internal nodes receiving 100% bootstrap support in all ML analyses regardless of the amount of missing data included. The 0% missing matrix (no type I missing data; supplementary table S1, Supplementary Material online) lacked the power to resolve nodes among genera in the gallopheasant complex (*Syrmaticus*, *Chrysolophus*, *Phasianus*, *Catraeus*, *Crossoptilon*, *Lophura*), whereas nodes relating these taxa were well supported in larger sequence alignments that included type I missing data. In general, node support increased with 1) alignment size and 2) the amount of missing data allowed from the 0% missing threshold to the 25% missing threshold. Nodal support in the 50% missing and total evidence matrices was largely similar to the 25% missing matrices. The observation that bootstrap support largely reached a point of diminishing returns with the 25% missing matrix was consistent with the limited increase in informative characters for the 50% missing and total evidence data matrices (table 1). The lone exception to this pattern was support for placement of *Pternistis ahantensis* (node Q; figs. 2 and 3), which received weak support in all analyses. We recovered highest support for node Q in the 5% missing matrix (52–57% bootstrap support, depending on settings); it received 32–39% bootstrap support in other matrices. Only 5 nodes failed to receive 100% bootstrap support in partitioned analysis of the 25% missing, 50% missing, and total evidence matrices for the 91 taxon, rooted, galliform ML tree (fig. 2).

In our RAxML analyses, neither the bootstrapping algorithm nor the approach used to accommodate among-sites rate heterogeneity (i.e., fast bootstrapping with GTRCAT vs. thorough bootstrapping with GTR+Γ) had a major effect on support values across the phylogeny (figs. 2 and 3). Likewise,

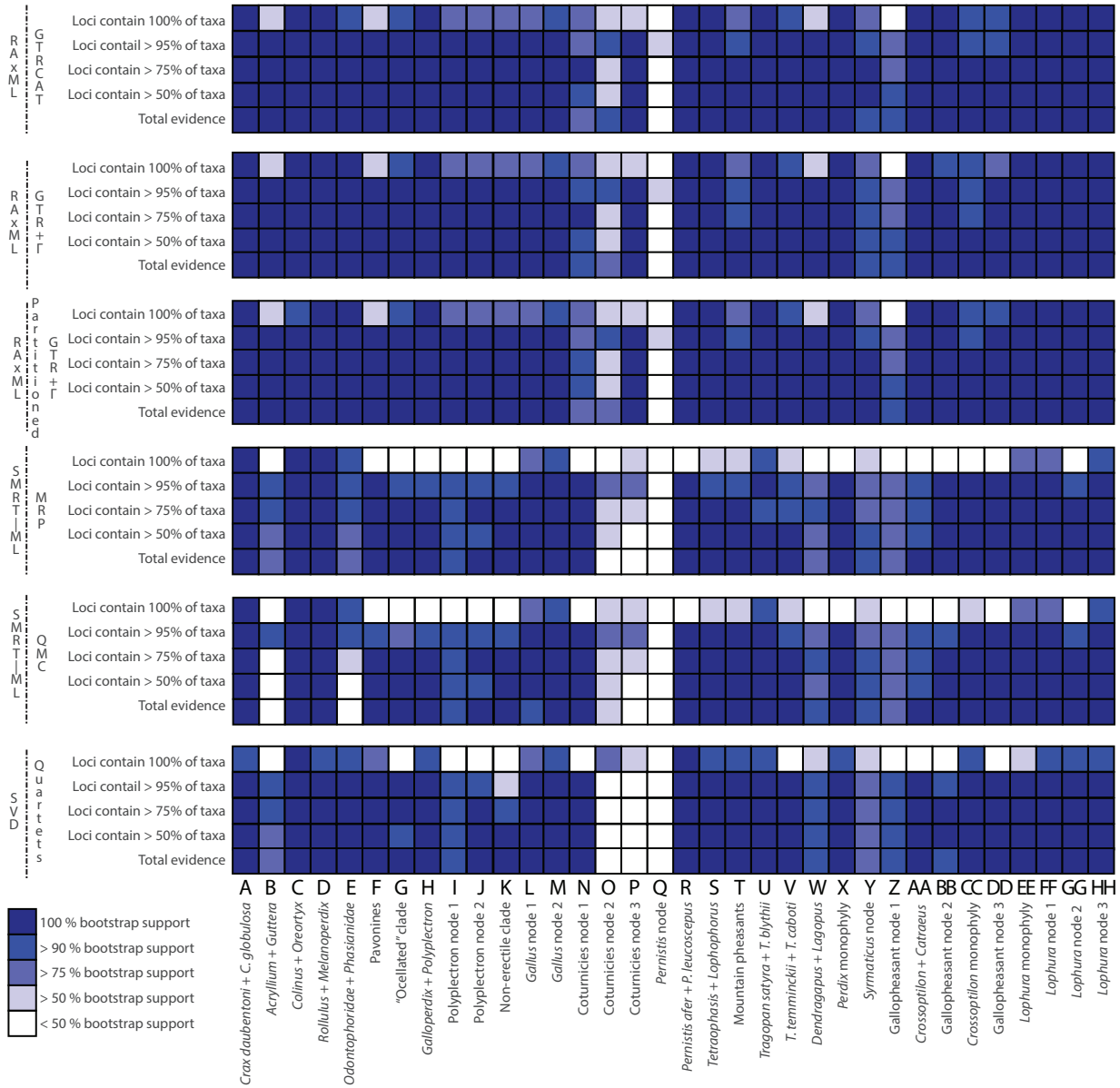

Fig. 3. Support values of selected nodes across analyses executed on concatenated matrices. Columns represent all nodes in the phylogeny with <100% bootstrap support for the 0% missing matrix, or where some gene tree reconciliation analyses conflicted with concatenation. Rows represent RAxML (rapid bootstrapping with GTRCAT, thorough bootstrapping with GTR+Γ unpartitioned, and thorough bootstrapping with GTR+Γ partitioned), SMRT-ML with MRP quartet reconciliation, SMRT-ML with QMC quartet reconciliation, and SVDquartets.

the support values were similar regardless of whether the analyses were partitioned or unpartitioned. For the small number of nodes with bootstrap support <100%, the bootstrap values we obtained with different settings were within a few percentage points of one another and there were no settings that consistently increased or decreased support across the phylogeny (fig. 3 and supplementary fig. S1, Supplementary Material online).

Quartet-based Phylogenetic Inference Consistent in the Anomaly Zone

SMRT-ML (DeGiorgio and Degnan 2010) and SVDquartets (Chifman and Kubatko 2014) produced phylogenies similar to standard concatenated ML with a few exceptions. In the SMRT-ML analysis of the smallest (i.e., 0% missing) matrix, two taxa, *Galloperdix spadicea* and *P. perdix*, having large proportions of type I and type II missing data (supplementary table S1, Supplementary Materials online) were placed in unexpected positions with low bootstrapping support. These placements conflicted with all ML trees, all SVDquartets trees, and all other SMRT-ML trees inferred from larger matrices, which placed *Galloperdix* sister to *Polyplectron* (node H) and recovered *Perdix* monophyly (node X; fig. 3 and supplementary figs. S2–S4, Supplementary Material online). Otherwise, SMRT-ML and SVDquartets were similar to concatenated ML [Robinson–Foulds (RF) distances generally <5; fig 2 inset and supplementary table S2, Supplementary Material online], except that they failed to recover sufficient bootstrap support to resolve relationships of *Alectoris* and *Ammoperdix* within the clade of Old World quail and relatives (nodes N, O, P; fig. 3).

In general, bootstrap support for SMRT-ML and SVDquartet phylogenies increased with matrix size, with the exception of nodes pertaining to *Guttera pucherani* (a low-yield sample with a large proportion of type I and type II missing data; table S1, Supplementary Materials online), which generally decreased in support as matrix size increased (fig. 3 and supplementary fig. S2–S4, Supplementary Material online). This was especially true with SMRT-ML using Quartet MaxCut QMC (QMC; supplementary fig. S3, Supplementary Materials online; Snir and Rao 2012), where *Guttera* was pulled outside of Numididae with moderate support in the 25% missing, 50% missing, and total evidence matrices. Otherwise, reconstructions of SMRT-ML quartets were qualitatively similar using QMC and Matrix Representation Parsimony (MRP). The 0% missing trees inferred with SMRT-ML and SVDquartets had considerably lower bootstrap support than those inferred with concatenated ML. Otherwise, bootstrap support values inferred with SVDquartets were slightly greater than those inferred with SMRT-ML, and slightly less than those inferred with concatenated ML. Like concatenated ML analyses, SMRT-ML and SVDquartets support values reached diminishing returns at $>$ 25% missing data.

Gene Tree Inference and Gene Tree Reconciliation Inference

Of 4,817 loci recovered in 4 or more galliform taxa, 4,613 contained at least one informative site and were used for downstream gene tree reconciliation in ASTRAL and ASTRID. For nine taxa (*Crax globulosa*, *Gu. pucherani*, *Melanoperdix niger*, *Tetraophasis obscurus*, *Tragopan satyra*, *Syrmaticus reevesii*, *Crossoptilon auritum*, *Lophura diardii*, and *Lophura leucomelanos*), ASTRAL phylogenies inferred using all gene trees (for all data completeness thresholds) consistently conflicted with those inferred using concatenated ML, SMRT-ML, and SVDquartets (RF distances 25–54), often with strong support (fig. 2–5). In each instance, trees inferred using ASTRAL placed these nine taxa toward the root when compared with their placement in concatenated ML and SMRT-ML topologies, creating a more pectinate tree shape (fig. 5 and supplementary fig. S5, Supplementary Material online). A review of N50s recovered for these nine taxa (supplementary table S1, Supplementary Material online) indicated that each had large amounts of type II missing data "within" loci (367, 334, 268, 303, 357, 348, 321, 350, and 367 bp N50, respectively) when compared with the entire data set (median 374 bp). Inferences using ASTRID were similar (fig. 5 and supplementary fig. S6, Supplementary Material online), except that 13 taxa with low N50s were placed toward the root with respect to those inferred with concatenated ML, SMRT-ML, and SVDquartets. These included the nine taxa affected in ASTRAL plus four additional taxa (*Arborophila rufogularis*, *Centrocercus urophasianus*, *Oreortyx pictus*, *Pternistis leucoscepus*; 266, 331, 348, and 342 bp N50 respectively) that were placed toward the root with respect to those inferred with concatenated ML, SMRT-ML, and SVDquartets.

In contrast, when we analyzed only more informative gene trees (i.e., those inferred from either the 25% or 5% most informative UCE loci) using ASTRAL, resulting topologies, including relationships for the nine problematic taxa, were similar to concatenated ML, SMRT-ML, and SVDquartets analyses with moderate to strong support (figs. 2–4; RF distances generally $<$10). Indeed, topologies inferred with ASTRAL using only the 5% most variable loci differed from the concatenated tree at only one node with strong bootstrap support: Placement of the peafowl clade (sister to the most recent common ancestor of *Polyplectron* and *Galloperdix* in concatenated ML, SMRT-ML, and SVDquartets; sister to the most recent common ancestor of *Gallus* and *Coturnix* in ASTRAL). The 0% missing data set was an exception to this pattern, and ASTRAL trees inferred from it were still largely unresolved. Unlike ASTRAL, ASTRID inferred unlikely relationships for several taxa with large proportions of type II missing data (e.g., *Gu. pucherani*, *Or. pictus*) even when only the most informative gene trees were used as input. Across all corresponding analyses, bootstrap support values inferred using ASTRID were considerably lower than those inferred with ASTRAL.

Discussion

Effects of Missing Data across Methods

Phylogenetic analysis of concatenated data reinforce the idea that increasing loci and nucleotides at the cost of increased missing data improves support at nodes that are difficult to reconstruct (Wagner et al. 2013; Huang and Knowles 2014; Streicher et al. 2016). However, few nodes in our galliform phylogeny showed increasing support when matrix size increased beyond the 25% missing threshold (an exception was node Z; figs. 2–4). Streicher et al. (2016) demonstrated a similar pattern in iguanian lizards, but in those analyses they observed diminishing returns at a 50% missing threshold. We expect that the point of diminishing returns differs between empirical data sets, and that data exploration by individual researchers is needed to determine appropriate thresholds. Diminishing returns may stem from the fact that as loci are added with increasing proportions of missing taxa/cells, they are less likely to include sequence data relevant to unresolved nodes in a phylogeny. Including taxa with relatively poor locus recovery (high type I missing data, taxa where we recovered approximately 25% of total loci) is also justified, and our results suggest that taxon exclusion to increase data completeness is likely to be unnecessarily cautious. In fact, there is evidence that increased taxon sampling can aid in resolving problems associated with long branch attraction even at the cost of increasing missing data (Wiens and Tiu 2012).

Similar to standard concatenated ML inference, support inferred with quartet methods (SMRT-ML and SVDquartets) improved with increasing numbers of loci. However, when compared with traditional concatenated ML, longer alignments were needed to produce strongly supported trees. For example, in the 0% missing concatenated ML tree, only two nodes had $<$50% bootstrap support (BS). In comparison, the 0% missing SVDquartets tree contained 5 such nodes and

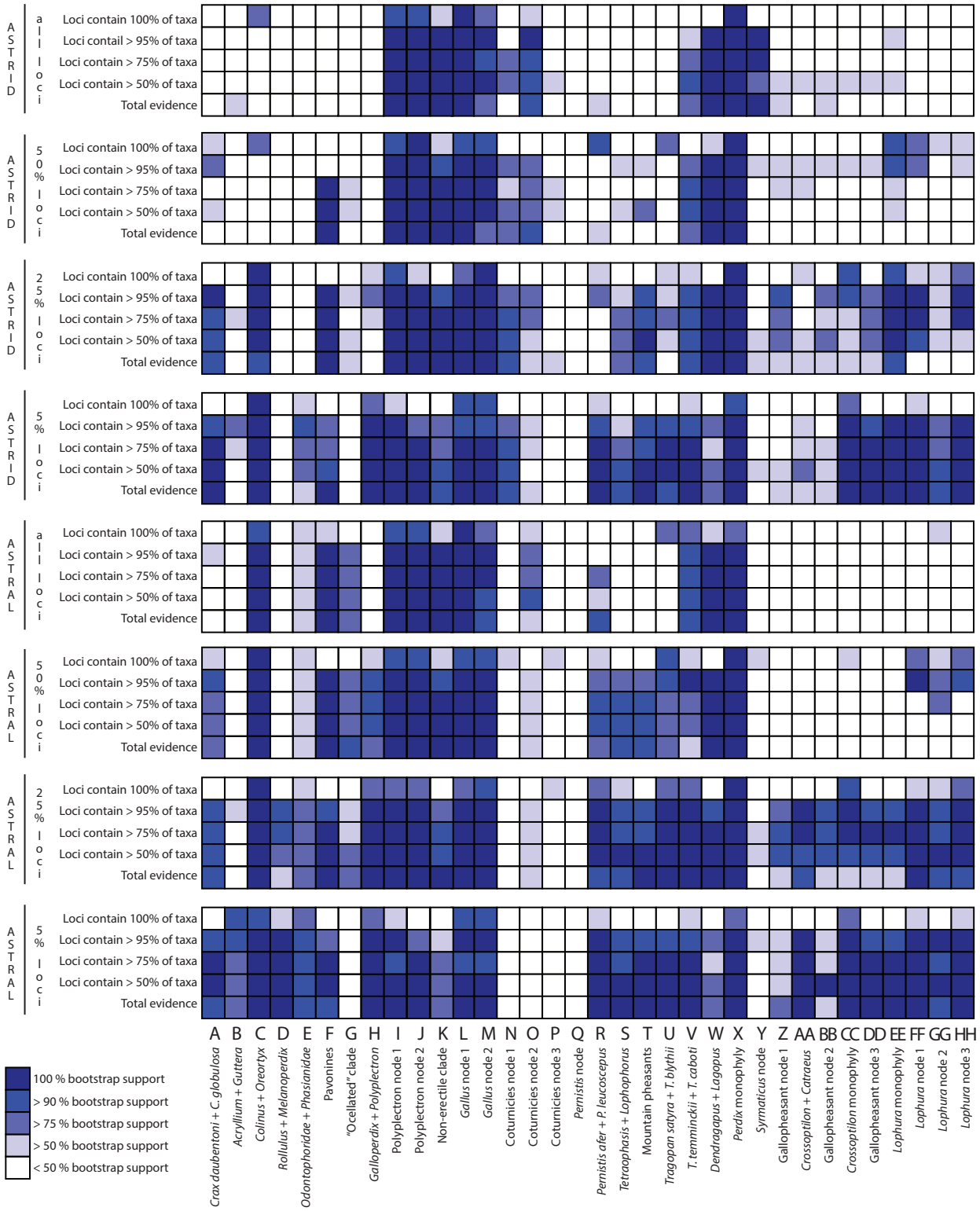

Fig. 4. Support values of selected nodes across gene tree reconciliation analyses. Columns represent all nodes in the phylogeny with <100% bootstrap support for the 0% missing matrix, or where some gene tree reconciliation analyses conflicted with concatenation. Rows represent ASTRID and ASTRAL gene tree reconciliation (with trees from all informative loci, 50% most informative loci, 25% most informative loci, and 5% most informative loci) analyses of the 0% missing (loci contain 100% of taxa), 5% missing (loci contain >95% of taxa), 25% missing (loci contain >75% of taxa), 50% missing (loci contain >50% of taxa), and the total evidence (all loci with >4 taxa) matrices.

the 0% missing SMRT-ML tree contained 19 such nodes. SMRT-ML and to a lesser extent SVDquartets appear to have reduced power to infer evolutionary relationships when compared with standard ML inference, as expected based on prior studies (DeGiorgio and Degnan 2010; Sun et al. 2014). Given sufficient data, however, these methods produced results similar to standard concatenation (figs. 2, inset, and 3).

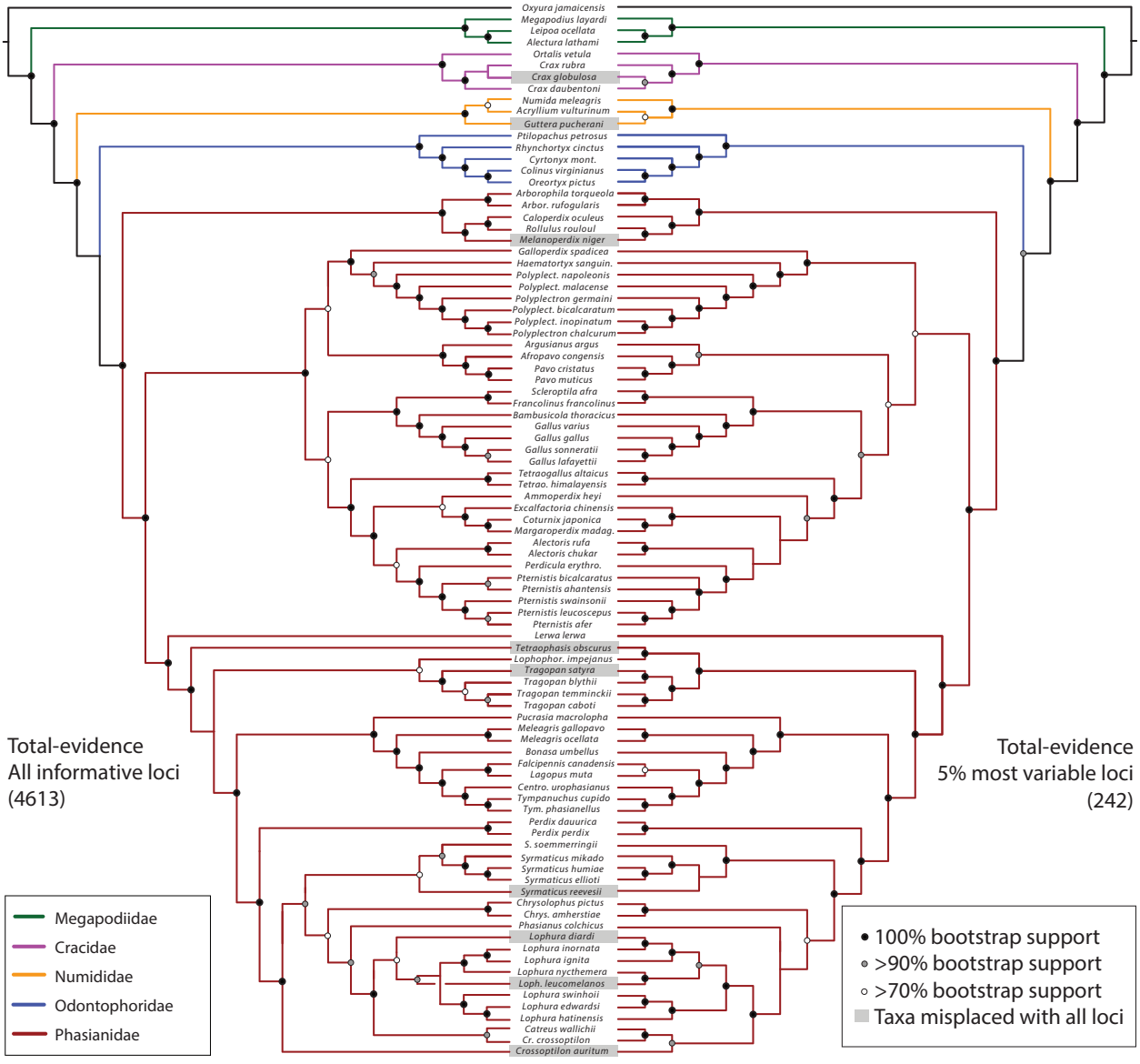

Fig. 5. Comparison of topology and support inferred with ASTRAL gene tree reconciliation. All (left) or only the most informative 5% of gene trees (right) are used as input. Nine taxa, highlighted in gray, are placed toward the root of the tree when all gene trees are used.

In contrast to results from concatenated ML, SMRT-ML, and SVDquartets analyses, gene tree reconciliation using ASTRAL inferred wildly different relationships among nine taxa (10% of included species). Gene tree reconciliation with ASTRID performed even more variably, with an additional four taxa placed in unexpected positions (14% of included species). These putatively spurious results do not appear to be related to increasing numbers of loci or type I missing data in alignments. Results for 5% missing, 25% missing, 50% missing, and total evidence matrices were all similar (the 0% missing had few resolved nodes, suggesting low power). Rather, the odd behavior using gene tree reconciliation approaches appeared to be related to the information content of the specific UCE loci we included (see below)—ASTRAL and ASTRID results improved markedly when we culled uninformative loci from the data set.

Uninformative Gene Trees, Biased Gene Trees, and Gene Tree Reconciliation

In contrast to analyses including all loci, using only the 25% most informative loci (ASTRAL; fig. 4) or the 5% most informative loci (ASTRAL and ASTRID; fig. 4) produced results largely congruent with those of ML, SMRT-ML, and SVDquartets inference. We observed this behavior in all alignments except the 100 complete matrix, which was still largely unresolved. Thus, inclusion of low information content loci in gene tree reconciliation approaches appears to hinder phylogenetic inference while also increasing computation time. Our gene tree reconciliation results clearly demonstrate that input gene tree resolution has a dramatic effect on inference.

Our results also suggest that the apparent poor performance of gene tree reconciliation is due to partial sequence data for certain taxa/loci (type II missing data) rather than

entire missing loci (type I missing data, which appeared to have little effect). Variation in enrichment efficiency, due to input sample quality or stochastic factors, along with stochastic variation in sequencing coverage, produces shorter contigs (low N50) for some taxa. Taxa with shorter contig lengths have fewer informative sites, such that estimates of gene trees with respect to those taxa will be poor (Simmons 2014) when compared with more data-rich samples, on average, across all loci. Thus, samples having shorter contigs are more likely to act as "unstable" or "rogue" taxa (Thomson and Shaffer 2010; Aberer et al. 2013; Goloboff and Szumik 2015) during individual gene tree inference. In the coalescent gene tree reconciliation analyses, erroneous placements of low N50 taxa in individual gene trees are attributed to ILS, rather than a systematic error introduced by missing sequence data. The net result is that low N50 taxa are systematically biased toward the root of the tree relative to their putatively true position (figs. 1 and 5). Examination of average contig lengths of these unstable taxa suggests that even relatively little missing data can cause this problem. For example, *Cra. globulosa* was unstable although its N50 of 367 bp was only 7 bp shorter than that of the mean N50 for the entire data set.

These results do not reflect theoretical problems of gene tree reconciliation analyses. Rather, they reflect a consequence of data collection procedures. Interestingly, not all taxa having low N50s suffered from this bias, and differences were apparent between reconciliation methods. For example, using ASTRAL, the taxon with the shortest N50 (*A. rufogularis* 266 bp), was always found sister to *Arborophila torquata* with strong support. With ASTRID, *A. rufogularis* was placed toward the root when all gene trees were analyzed, but ASTRID recovered *Arborophila* monophyly when only more informative gene trees were input. Clearly, N50 and reconciliation method are not the only factors affecting putatively erroneous placement of taxa: The relative branch lengths are also important. The branch joining *Arborophila* taxa to other Galliformes is relatively long, with ample time for substitutions to accrue; thus gene trees were reconstructed fairly reliably despite a low N50 in *A. rufogularis*. In general, both number of misplaced taxa and overall support values suggest that ASTRID is more sensitive to type II missing data than ASTRAL.

Limiting the set of input trees for gene tree reconciliation analyses to those based on more information-rich loci, from which gene trees can be estimated more reliably, provides a potential solution to the problem of biased gene trees without requiring the removal of all low N50 taxa from analyses. In our analyses, eliminating relatively uninformative gene trees greatly improved the apparent performance of gene tree reconciliation. In light of this observation, revisiting previous studies that employed gene tree reconciliation on low-variation markers (McCormack et al. 2013; Streicher et al. 2016) may be warranted to determine if selection of more informative gene trees can improve support. Another suitable approach would be to limit missing data within loci by trimming contigs to the shortest sequences. However, this approach would work poorly with our data set: Most UCE variation occurs in flanking regions at the 5′ and 3′ ends of sequences, and trimming to the shortest contigs leaves only the largely invariant core UCE region for analysis. Even in the most variable UCE loci that contain hundreds of informative sites across Galliformes, such extreme truncation would leave only a handful of informative sites, producing only poorly resolved gene trees for downstream gene tree reconciliation.

Our results indicating that gene tree reconciliation analyses perform better when uninformative gene trees are excluded raises a key question: How should researchers decide which gene trees to include? As a first step, our study chose arbitrary thresholds and compared performances. Although results from these arbitrary thresholds were qualitatively similar, there were slight differences in inferred topologies and estimated support. For example, the 25% most informative loci and the 5% most informative loci differed in their placement of *Pavo* and its relatives, one of the most challenging nodes to resolve in galliforms (Sun et al. 2014). Thus, user decisions regarding locus inclusion can influence results (Betancur-R et al. 2014). Nonetheless, other analyses have suggested that UCE flanking regions have excellent phylogenetic informativeness (Gilbert et al. 2015). Although the best metric for phylogenetic informativeness and the appropriate thresholds for inclusion loci in gene tree reconciliation analyses remain unclear, we found that for UCE loci, a commonly used metric (Townsend 2007; López-Giráldez and Townsend 2011) was strongly correlated with the number of parsimony informative sites (supplementary fig. S7, Supplementary Material online). This prompted us to use the simpler metric. Ideally, a multilocus approach using methods developed to identify and prune rogue taxa in supermatrices (Thomson and Shaffer 2010; Aberer et al. 2013; Goloboff and Szumik 2015) could objectively identify loci and/or taxa to be pruned from data sets.

An alternative and perhaps more robust solution would be to reduce dependence on gene tree reconciliation for phylogenomic inference when there are concerns about gene tree reliability. Methods like SMRT-ML and SVDquartets are executed on concatenated sequences, avoiding problems with type II missing data, biased or poor estimation of gene trees, or user choice of input gene trees. These methods also scale well to large data sets, and they are likely to scale to very large data matrices if it is possible to sample subsets of quartets without sacrificing accurately. We recommend further testing of these methods to determine their robustness to the many ways that empirical data may violate their assumptions.

Given the numerous rapid radiations at various depths within Galliformes (Kimball and Braun 2014), one might have expected different results inferred from concatenated versus coalescent approaches. Instead, we found that results from each framework were largely congruent, although in the case of gene tree reconciliation methods this required limiting the input to the more informative gene trees. The observed congruence of estimate of the species tree obtained using concatenated and coalescent methods suggests that the primary cause of gene tree discordance in our Galliformes UCE data set was error in gene tree estimation due to character limitation rather than ILS per se. No doubt some discordance among true gene trees reflects ILS, but

congruence between approaches suggests that ILS is relatively low in the galliform tree (Mirarab, Bayzid, and Warnow 2014), and that our ML estimate of galliform phylogeny was not affected by the anomaly zone. We also note that there may be additional sources of discordance among true gene trees for the galliforms due to other processes, such as lateral gene transfer (hybridization is considered to be common in Galliformes; Johnsgard 1970, 1988; Dong et al. 2013). However, spurious results obtained from including all gene trees in ASTRAL and ASTRID analyses reinforced the idea that error in gene tree estimation may often be the most important source of error in phylogenetic inference (Patel et al. 2013; Gatesy and Springer 2014). Thus, error in gene tree estimation should be considered when weighing selection of concatenated and coalescent approaches for low-variation markers like UCEs.

There are two clades in the galliform tree where concatenated ML differed from coalescent analyses with strong support. First, concatenated ML differed from gene tree reconciliation (when only the 5% most informative gene trees were considered) with respect to placement of the peafowl clade (node G; figs. 2–5). If this difference is attributable to a bias in phylogenetic reconstruction by standard ML analyses of concatenated data, we would predict SMRT-ML and SVDquartets to agree with gene tree reconciliation. Yet, our SMRT-ML and SVDquartets results agree with standard concatenated ML. Curiously, ASTRAL results (when the 25% and 50% most informative gene trees were considered) also agree with concatenated ML, SMRT-ML, and SVDquartets, suggesting that user choice of input trees has a strong effect on node G. Second, standard concatenated ML recovered strong support for placement of *Ammoperdix* and *Alectoris* (nodes O and P), but coalescent analyses, including SMRT-ML and SVDquartets, found little support for any relationships pertaining to these nodes. There are two possible interpretations for this result. Concatenated ML is thought to have greater power than coalescent methods to identify relationships when ILS is low. If this is the case, then the inferred concatenated ML tree may be reliable. However, if there is substantial ILS with respect to nodes O and P, concatenated ML could be positively misleading. The first of these two hypotheses seems more likely given the topology and branch lengths, because the subtree defined by nodes O and P is maximally asymmetric and anomaly zone problems reflect the higher probability of symmetric gene trees given asymmetric species trees (Degnan and Rosenberg 2006; Rosenberg 2013), and the recovery of asymmetric trees in concatenated analyses has been used to argue against the existence of a bias due to the anomaly zone (Harshman et al. 2008; Smith et al. 2013). Further exploration of these nodes using more rapidly evolving markers and/or increased taxon sampling would be ideal to resolve the relationship nodes defined O and P with greater confidence.

Data Quality from Historical Samples

One objective of this study was to explore the potential of using historical DNA extracted from museum skin toepads (Mundy et al. 1997) as source material for UCE enrichment and sequencing (McCormack et al. 2015). Use of historical material (often referred to as "ancient DNA") is common in systematics, and its use has the clear benefit of allowing researchers to increase taxon sampling when fresh tissues are unavailable (McCallum et al. 2013; Heupink et al. 2014; Mitchell et al. 2014). However, the fragmented nature of historical DNA can lead to polymerase chain reaction (PCR) contamination, PCR errors (Sefc et al. 2007), and preferred amplification of pseudogenes (Greenwood et al. 1999), which compromises phylogenetic inference and may often go unnoticed for years (Zuccon and Ericson 2010; Moyle et al. 2013, 2015). The challenge of producing extensive character data sets with historical DNA has limited researchers to (often partial) sampling of one or a few loci, frequently just the mitochondrion, which may also limit or mislead phylogenetic inference (Maddison 1997).

Consistent with previous studies that have advocated the use of historical DNA in massively parallel sequencing (Knapp and Hofreiter 2010; Mason et al. 2011; McCormack et al. 2015), our results demonstrate that target capture of historical DNA can be successful with little alteration of protocols, and that inclusion of historical samples gives a clearer picture of galliform phylogeny (Sun et al. 2014). Yet, uncritical use of these samples, which have a lower N50 than fresh tissues (supplementary table S1, Supplementary Material online), could give rise to spurious results when using gene tree reconciliation frameworks (figs. 3 and 4). Thus, historical samples are an important source for target capture, but they are better utilized minimally rather than as a routine substitute for high-quality source material.

Toward a Robust Phylogeny of Galliformes

We recovered congruent and strongly supported topologies using numerous phylogenomic inference strategies, each subject to different limitations and biases. Thus, UCE results finally clarify many historical relationships that have remained problematic. For example, relationships within the junglefowl (*Gallus*), which includes the domestic chicken, have been weakly supported and variable in topology (Kimball and Braun 2008, 2014; Wang et al. 2013), although results using whole mitochondrial genomes strongly support placing *Ga. gallus* sister to *Gallus varius* (fig. 6; Meiklejohn et al. 2014). In contrast to those results from mitochondrial data, our concatenated and coalescent results unequivocally support placing *Ga. gallus* sister to a *Gallus sonneratii* + *Gallus lafayettii* clade (figs. 2–6). Given the contrast to strong supported in mitogenomic studies, this could either indicate cyto-nuclear discordance in this clade of closely related birds, or a rooting problem in the mitogenomic tree.

Relationships among genera of Old World quails and relatives have also been largely unresolved in previous studies (reviewed by Crowe et al. 2006; Wang et al. 2013), although only a few of these genera have been included in more than one or two studies. Concatenated ML analyses inferred a well-supported multilocus hypothesis of this group, whereas

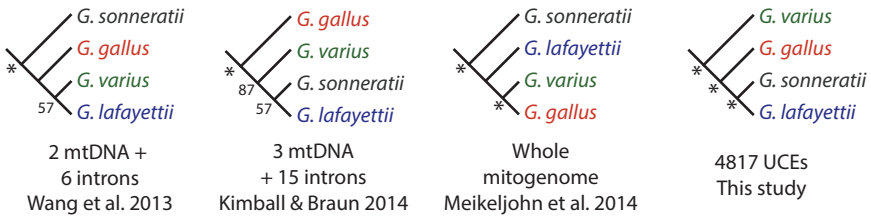

Fig. 6. Recent phylogenetic hypotheses for the junglefowl genus *Gallus*. Node labels refer to ML bootstrap percentages, with * = 100%. Previous studies recovered differing topologies, often without strong support. All UCE analysis, including ASTAL runs using all gene trees (which performed poorly with respect to many taxa), recovered this topology with strong support.

SMRT-ML, SVDquartets, and gene tree reconciliation analyses failed to resolve polytomies among these genera (figs. 3–5). In contrast to Crowe et al. (2006), who found subcontinental *Perdicula* sister to African *Pternistis* + *Ammoperdix*, we identified sister relationships between *Pternistis*/*Perdicula* and *Ammoperdix*/Old World quails (*Coturnix*/*Excalfactoria*/*Margaroperdix*).

Two genera that have not been included in previous molecular analyses of Galliformes (*Lerwa*, *Melanoperdix*; each sampled with historical DNA from museum skin toepads) were placed with confidence in this study (figs. 2–5). Previous authors have expressed uncertainty over evolutionary relationships of the Sino-Himalayan *Lerwa lerwa*, and suggested *Tetraogallus* as a possible relative based on gross morphology (Johnsgard 1988), or *Ithaginis* based on plumage characters of downy chicks (Potapov 2000). Our results support that *L. lerwa* is a unique evolutionary lineage sister to the "erectile clade" of pheasants, grouse, and partridges (Kimball et al. 2008). Previous analyses have also placed *Ithaginis* sister to the erectile clade (Wang et al. 2013; Meiklejohn et al. 2014). Inclusion of both *Ithaginis* (not available for this study) and *Lerwa* is needed to determine their relative relationships to the erectile clade. *Melanoperdix niger* was recovered sister to *Rollulus rouloul*, with which it is sympatric over its entire distribution. These taxa are placed in a clade of Southeast Asian forest-dwelling partridges, a result previously suggested by similarities in female plumage and gross morphology (Johnsgard 1988).

Overall, phylogenetic analyses of UCEs were successful in resolving problematic nodes in the galliform tree of life highlighted by previous studies, providing further evidence that UCEs are suitable markers for resolving some of the most challenging empirical problems in molecular systematics. Furthermore, where taxon membership has overlapped, UCEs analyzed here inferred a well-resolved galliform backbone topology identical to that inferred with other marker classes such as UTRs (Bonilla et al. 2010), nuclear introns (Hackett et al. 2008), and other conserved genomic regions (Prum et al. 2015). However, care needs to be taken in analyses, particularly gene tree reconciliation, when including poorly resolved gene trees. This suggests that building robust large-scale phylogenies for other challenging-to-resolve groups is within reach, and that future comparative studies using these phylogenies will produce robust insight into macroevolutionary processes.

Materials and Methods

Sequence Capture, Assembly, and Alignment of UCEs

We selected 86 galliform taxa and one outgroup (Anseriformes: *Oxyura jamaicensis*) for UCE enrichment, including all major clades identified by previous studies. Taxa were selected to 1) cover the breadth of galliform diversity; 2) focus on problematic nodes, particularly surrounding model taxa in the Phasianidae; and 3) identify relationships of two enigmatic, monotypic genera not included in previous molecular phylogenetic studies due to a lack of fresh tissue resources (*Me. niger*, *L. lerwa*). We extracted genomic DNA from fresh tissues (most samples) or toepads of museum specimens; sample information can be found in supplementary table S1, Supplementary Material online.

Sequence capture libraries were prepared using an approach modified from Faircloth et al. (2012). We prepared Nextera sequencing libraries using the manufacturer's protocol (Illumina, Inc., San Diego, CA), but using primers with custom index tags (Faircloth and Glenn 2012). We pooled 8 samples together, and enriched each library pool for 5,060 UCE loci targeted by 5,472 probes (Mycroarray, Ann Arbor, MI; http://www.mycroarray.com/mybaits/mybaits-UCEs.html). Enriched libraries were amplified with 18 PCR cycles, quantified using qPCR (quantitative PCR; Kapa Biosystems), and sequenced in a single Illumina HiSeq 2000 Lane (75 nt paired-end reads; UC Irvine Genomics High-Throughput Facility). DNA extracts from toepads underwent the same library preparation procedures as fresh tissue.

We assembled quality-controlled reads into contigs de novo using Trinity (Grabherr et al. 2011), and added UCE loci of *Ga. gallus*, *M. gallopavo*, *Cot. japonica*, and *Col. virginianus* extracted from published genome assemblies (International Chicken Genome Sequencing Consortium 2004; Dalloul et al. 2010; Kawahara-Miki et al. 2013; Halley et al. 2014). Sequences for UCE loci obtained from four or more taxa (n = 4,817) were aligned using MAFFT 7 (Katoh et al. 2002; Katoh and Standley 2013). We trimmed ends of alignments when 35% of cells were missing. Resulting final taxon sampling included 90 taxa and 89 of 298 currently recognized galliform species (Gill and Donsker 2015).

Alignment Filtering Matrix Construction

We constructed sets of UCE alignments with five thresholds of taxonomic completeness (table 1) with the PHYLUCE 1.5 pipeline (Faircloth et al. 2012, Faircloth 2015). Alignments

included 1) all taxa for each UCE locus (0% missing), 2) greater than 95% of taxa present for each UCE locus (5% missing), 3) greater than 75% of taxa present for each UCE locus (25% missing), 4) greater than 50% of taxa present for each UCE locus (50% missing), and 5) total evidence, which included all 4,817 UCE loci recovered in 4 or more taxa. We used each of these five sets of alignments in downstream phylogenetic inference with concatenated ML, concatenated quartet-based, and gene tree reconciliation approaches.

Concatenated Maximum Likelihood Phylogenetic Inference

For each concatenated alignment (0% missing, 5% missing, 25% missing, 50% missing, total evidence), we implemented PartitionFinder 1.1 (Lanfear et al. 2012) using the "hcluster" algorithm (using default weighting: rate = 1, base = 0, model = 0, alpha = 0) and the GTR+Γ model of sequence evolution with each UCE locus as a data subset. We conducted ML phylogenetic inference in RAxML 8.1.1 (Stamatakis et al. 2008; Stamatakis 2014). For each of the 5 concatenated alignments, we computed an ML tree and 500 bootstrap replicates with 1) rapid nonparametric bootstrapping of the unpartitioned data set using the GTRCAT approximation, 2) thorough nonparametric bootstrapping of the unpartitioned data set using GTR+Γ, and 3) thorough nonparametric bootstrapping of the partitioned data set using GTR+Γ. All phylogenetic analyses were computed at the University of Florida High-Performance Computing Center.

Quartet-based Phylogenetic Inference Consistent in the Anomaly Zone

We estimated phylogenies for each of the five concatenated matrices (0% missing, 5% missing, 25% missing, 50% missing, total evidence) using two quartet-based methods that take advantage of the fact that there are no anomalous trees for a rooted quartet. These approaches are executed on concatenated matrices, and therefore may be less sensitive to missing data than gene tree reconciliation. First, we implemented a modified supermatrix rooted triples approach (SMRT-ML; DeGiorgio and Degnan 2010). All supermatrix-rooted triples (effectively all possible quartets that contain the outgroup *Ox. jamaicensis*) were inferred with RAxML 8.1.1 under the GTR+Γ model from unpartitioned concatenated matrices using a custom Perl script. Phylogenies were constructed from SMRT-ML quartets were using two methods: 1) We built MRP matrices from quartets with Clann (Creevey and McInerney 2005), which then were executed in PAUP* 4.0b10 (Swofford 2003), and 2) we reconstructed phylogenies from quartets directly with QMC 3.0 (Snir and Rao 2010, 2012). We produced 100 nonparametric SMRT-ML bootstraps for each of the 5 data matrices. Second, we implemented SVDquartets (Chifman and Kubatko 2014), a method that infers quartets based on summaries of SNPs in a concatenated sequence matrix. We invoked SVDquartets in PAUP* 4.1a146, sampling all quartets, and we constructed phylogenies using QMC 3.0 (Avni et al. 2015). We inferred SVDquartets for 100 nonparametric bootstraps for each of the 5 data matrices.

Multispecies Coalescent Inference Using Gene Tree Reconciliation

We inferred phylogenies for each set of five sets of alignments (0% missing, 5% missing, 25% missing, 50% missing, total evidence) under the multispecies coalescent using two gene tree reconciliation algorithms, ASTRAL 4.4.4 (Mirarab, Reaz, et al. 2014) and ASTRID (Vachaspati and Warnow 2015). ASTRAL (Mirarab, Reaz, et al. 2014) takes sets of gene trees (or bootstraps of unrooted gene trees) and computes the phylogeny which agrees with the largest number of quartet trees induced by the gene tree set. ASTRID takes sets of gene trees and computes the phylogeny from internode distance.

ASTRAL and ASTRID use unrooted gene trees and allow for missing taxa, therefore allowing inclusion of loci for which the outgroup taxon is missing. Note that a rewritten version of ASTRAL (ASTRAL II; Mirarab and Warnow 2015) handles missing data differently, and may outperform the standard version. However, like other coalescent-based gene tree reconciliation methods, ASTRAL and ASTRID assume that the input gene trees are estimated without error. To understand the effects of including uninformative loci that may bias gene tree reconciliation, we implemented a series of ASTRAL and ASTRID runs using 1) all gene trees, 2) gene trees from the 50% most parsimony informative loci, 3) gene trees from the 25% most parsimony informative loci, and 4) gene trees from the 5% most parsimony informative loci. To ensure that the number of parsimony informative sites is an accurate indicator of informativeness, we also calculated phylogenetic informativeness using PhyDesign (López-Giráldez and Townsend 2011) for a subset of loci with little missing data (supplementary fig. S7, Supplementary Material online). Because the method of Townsend (2007) was highly correlated with parsimony informativeness (supplementary fig. S7, Supplementary Material online), we selected gene trees for inclusion using only number of parsimony informative sites. To estimate gene trees, we computed 100 thorough bootstraps under GTR+Γ for all 4,638 loci containing 4 or more taxa and at least one informative site. We computed all four thresholds of locus variability (all gene trees, gene trees from the 50% most parsimony informative loci, gene trees from the 25% most parsimony informative loci, and gene trees from the 5% most parsimony informative loci) with each of our five sets of alignments of varying completeness (0% missing, 5% missing, 25% missing, 50% missing, total evidence).

Topological Comparisons

To compare phylogenetic results from different analyses with different thresholds of matrix completeness, we computed pairwise RF distances between majority rule consensus trees in PAUP* 4.0b10 (Swofford 2003), and visualized tree space using multidimensional scaling (Hillis et al. 2005) of the pairwise RF tree distance matrix computed in R 3.1 (R Core Team 2014).

Supplementary Material

Supplementary figures S1–S7 and tables S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank institutions, collectors, and collection managers who provided tissues for this study: The Field Museum of Natural History, the Florida Museum of Natural History, the Louisiana State Museum of Natural Science, and the University of Washington Burke Museum. Ricardo Betancur-R and two anonymous reviewers provided constructive comments to improve the manuscript. This work was supported by the National Science Foundation [grant numbers DEB-1118823 to (R.T.K. and E.L.B.) and DEB-1242260 (to B.C.F.)]. Alignments and treefiles are deposited at FigShare at http://figshare.com/account/projects/5596. Custom Perl scripts are available at GitHub at https://github.com/ebraun68/SMRT-ML. DNA Sequence read data are archived on NCBI SRA (PRJNA303085).

References

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning rogue taxa improves phylogenetic accuracy: an efficient algorithm and webservice. *Syst Biol*. 62:162–166.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. *Science* 304:1321–1325.

Betancur-R R, Naylor GJP, Ortí G. 2014. Conserved genes, sampling error, and phylogenomic inference. *Syst Biol*. 63:257–262.

BirdLife International. 2012. IUCN Red List for birds. Available from: http://www.iucnredlist.org/.

Bonilla AJ, Braun EL, Kimball RT. 2010. Comparative molecular evolution and phylogenetic utility of 3′-UTRs and introns in Galliformes. *Mol Phylogent Evol*. 56:536–542.

Braun EL, Kimball RT. 2001. Polytomies, the power of phylogenetic inference, and the stochastic nature of molecular evolution: a comment on Walsh et al. (1999). *Evolution* 55:1261–1263.

Burleigh JG, Kimball RT, Braun EL. 2015. Building the avian tree of life using a large-scale, sparse supermatrix. *Mol Phylogent Evol*. 84:53–63.

Chifman J, Kubatko L. 2014. Quartet inference from SNP data under the coalescent model. *Bioinformatics* 30:3317–3324.

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. 2012. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol Lett*. 8:783–786.

Crawford NG, Parham JF, Sellas AB, Faircloth BC, Glenn TC, Papenfuss TJ, Henderson JB, Hansen MH, Simison WB. 2015. A phylogenomic analysis of turtles. *Mol Phylogent Evol*. 83:250–257.

Creevey CJ, McInerney JO. 2005. Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* 21:390–392.

Crowe T, Bowie R, Bloomer P, Mandiwana T, Hedderson T, Randi E, Pereira S, Wakeling J. 2006. Phylogenetics, biogeography and classification of, and character evolution in, gamebirds (Aves: Galliformes): effects of character exclusion, data partitioning and missing data. *Cladistics* 22:495–532.

Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Ann Blomberg L, Bouffard P, Burt DW, Crasta O, Crooijmans RPMA, et al. 2010. Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis. *PLoS Biol*. 8:e1000475.

DeGiorgio M, Degnan JH. 2010. Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol Biol Evol*. 27:552–569.

DeGiorgio M, Syring J, Eckert AJ, Liston A, Cronn R, Neale DB, Rosenberg NA. 2014. An empirical evaluation of two-stage species tree inference strategies using a multilocus dataset from North American pines. *BMC Evol Biol*. 14:1–10.

Degnan JH. 2013. Anomalous unrooted gene trees. *Syst Biol*. 62:574–590.

Degnan JH, Rosenberg NA. 2006. Discordance of species trees with their most likely gene trees. *PLoS Genet*. 2:e68.

Degnan JH, Rosenberg NA. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol Evol*. 24:332–340.

Dong L, Heckel G, Liang W, Zhang Y. 2013. Phylogeography of Silver Pheasant (*Lophura nycthemera* L.) across China: aggregate effects of refugia, introgression and riverine barriers. *Mol Ecol*. 22:3376–3390.

Edwards SV. 2009. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–19.

Edwards SV, Liu L, Pearl D. 2007. High-resolution species trees without concatenation. *Proc Natl Acad Sci U S A* 104:5936–5941.

Faircloth BC. 2015. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*.

Faircloth BC, Glenn TC. 2012. Not all sequence tags are created equal: designing and validating sequence identification tags robust to indels. *PLoS One*. 8:e42543.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol*. 61:717–726.

Faircloth BC, Sorenson L, Santini F, Alfaro ME. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS One* 8:e65923.

Gatesy J, Springer MS. 2014. Phylogenetic analysis at deep timescales: unreliable gene trees, bypassed hidden support, and the coalescence/concatalescence conundrum. *Mol Phylogent Evol*. 80:231–266.

Gilbert PS, Chang J, Pan C, Sobel EM, Sinsheimer JS, Faircloth BC, Alfaro ME. 2015. Molecular phylogenetics and evolution. *Mol Phylogent Evol*. 92:140–146.

Gill F, Donsker D. 2015. IOC world bird names (version 5.3). Available from: http://www. worldbirdnames.org.

Goloboff PA, Szumik CA. 2015. Identifying unstable taxa: efficient implementation of triplet-based measures of stability, and comparison with Phyutility and RogueNaRok. *Mol Phylogent Evol*. 88:93–104.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 29:644–652.

Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutiérrez S, Castoe TA, et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* 346:1254449.

Greenwood AD, Capelli C, Possnert G, Pääbo S. 1999. Nuclear DNA sequences from late *Pleistocene megafauna*. *Mol Biol Evol*. 16:1466–1473.

Hackett SJ, Kimball RT, Reddy S, Bowie RCK, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, et al. 2008. A phylogenomic study of birds reveals their evolutionary history. *Science* 320:1763–1768.

Halley YA, Dowd SE, Decker JE, Seabury PM, Bhattarai E, Johnson CD, Rollins D, Tizard IR, Brightsmith DJ, Peterson MJ, et al. 2014. A draft de novo genome assembly for the Northern Bobwhite (*Colinus virginianus*) reveals evidence for a rapid decline in effective population size beginning in the Late Pleistocene. *PLoS One* 9:e90240.

Harshman J, Braun EL, Braun MJ, Huddleston CJ, Bowie RC, Chojnowski JL, Hackett SJ, Han KL, Kimball RT, Marks BD. 2008. Phylogenomic evidence for multiple losses of flight in ratite birds. *Proc Natl Acad Sci U S A*. 105:13462–13467.

Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol Biol Evol*. 27:570–580.

Heupink TH, van Grouw H, Lambert DM. 2014. The mysterious Spotted Green Pigeon and its relation to the Dodo and its kindred. *BMC Evol Biol*. 14:1–6.

Hillis DM, Heath TA, St. John. 2005. Analysis and visualization of tree space. *Syst Biol*. 54:471–482.

Hirst CE, Marcelle C. 2015. The avian embryo as a model system for skeletal myogenesis. *Results Probl Cell Differ*. 56:99–122.

Huang H, Knowles LL. 2014. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst Biol*.
International Chicken Genome Sequencing Consortium 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432:695–716.
Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346:1320–1331.
Johnsgard PA. 1970. A summary of intergeneric New World Quail hybrids, and a new intergeneric hybrid combination. *Condor* 72:85–88.
Johnsgard PA. 1988. The quails, partridges, and francolins of the world. London: Oxford University Press.
Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*. 30:3059–3066.
Katoh K, Standley DM. 2013. MAFFT Multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30:772–780.
Kawahara-Miki R, Sano S, Nunome M, Shimmura T, Kuwayama T, Takahashi S, Kawashima T, Matsuda Y, Yoshimura T, Kono T. 2013. Next-generation sequencing reveals genomic features in the Japanese quail. *Genomics* 101:345–353.
Kimball RT, Braun EL. 2008. A multigene phylogeny of Galliformes supports a single origin of erectile ability in non-feathered facial traits. *J Avian Biol*. 39:438–445.
Kimball RT, Braun EL. 2014. Does more sequence data improve estimates of galliform phylogeny? Analyses of a rapid radiation using a complete data matrix. *PeerJ* 2:e361.
Kimball RT, Braun EL, Ligon JD, Lucchini V, Randi E. 2008. A molecular phylogeny of the peacock-pheasants (Galliformes: *Polyplectron* spp.) indicates loss and reduction of ornamental traits and display behaviours. *Biol J Linnean Soc*. 73:187–198.
Kimball RT, Mary CMS, Braun EL. 2011. A macroevolutionary perspective on multiple sexual traits in the Phasianidae (Galliformes). *Int J Evol Biol*. 2011:1–16.
Kimball RT, Wang N, Heimer-McGinn V, Ferguson C, Braun EL. 2013. Identifying localized biases in large datasets: a case study using the avian tree of life. *Mol Phylogent Evol*. 69:1021–1032.
Knapp M, Hofreiter M. 2010. Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes* 1:227–243.
Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Syst Biol*. 56:17–24.
Kumar S, Filipski AJ, Battistuzzi FU, Kosakovsky Pond SL, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol*. 29:457–472.
Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol Biol Evol*. 29:1695–1701.
Le Douarin NM, Dieterlen-Lièvre F. 2012. How studies on the avian embryo have opened new avenues in the understanding of development: a view about the neural and hematopoietic systems. *Dev Growth Differ*. 55:1–14.
Leaché AD, Rannala B. 2011. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol*. 60:126–137.
Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*. 61:727–744.
Lemmon EM, Lemmon AR. 2013. High-throughput genomic data in systematics and phylogenetics. *Ann Rev Ecol Syst*. 44:99–121.
Liu L, Edwards SV. 2009. Phylogenetic analysis in the anomaly zone. *Syst Biol*. 58:452–460.
Liu L, Xi Z, Wu S, Davis CC, Edwards SV. 2015. Estimating phylogenetic trees from genome-scale data. *Ann N Y Acad Sci*. 1360:36–53.
López-Giráldez F, Townsend JP. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. *BMC Evol Biol*. 11:152
Maddison W. 1997. Gene trees in species trees. *Syst Biol*. 46:523–536.
Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ. 2010. Target-enrichment strategies for next-generation sequencing. *Nat Meth*. 7:111–118.
Mason VC, Li G, Helgen KM, Murphy WJ. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. *Genome Res*. 21:1695–1704.
McCallum J, Hall S, Lissone I, Anderson J, Huynen L, Lambert DM. 2013. Highly informative ancient DNA "Snippets" for New Zealand Moa. *PLoS One* 8:e50732.
McCormack JE, Faircloth BC. 2013. Next-generation phylogenetics takes root. *Mol Ecol*. 22:19–21.
McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res*. 22:746–754.
McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. *PLoS One* 8:e54848.
McCormack JE, Tsai WLE, Faircloth BC. 2015. Sequence capture of ultraconserved elements from bird museum specimens. *Mol Ecol Res*.
Meiklejohn KA, Danielson MJ, Faircloth BC, Glenn TC, Braun EL, Kimball RT. 2014. Incongruence among different mitochondrial regions: a case study using complete mitogenomes. *Mol Phylogent Evol*. 78:314–323.
Mirarab S, Bayzid MS, Boussau B, Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346:1250463–1250463.
Mirarab S, Bayzid MS, Warnow T. 2014. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. *Syst Biol*.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.
Mirarab S, Warnow T. 2015. ASTRAL-II: Coalescent -based species tree estimation with many hundreds of taxa and thousands of genes. *Bioinformatics* 31:i44–i52.
Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346:763–767.
Mitchell KJ, Llamas B, Soubrier J, Rawlence NJ, Worthy TH, Wood J, Lee MSY, Cooper A. 2014. Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution. *Science* 344:898–900.
Moyle RG, Filardi CE, Smith CE, Diamond J. 2009. Explosive Pleistocene diversification and hemispheric expansion of a "great speciator". *Proc Natl Acad Sci U S A* 106:1863–1868.
Moyle RG, Hosner PA, Jones AW, Outlaw DC. 2015. Phylogeny and biogeography of *Ficedula* flycatchers (Aves: Muscicapidae): novel results from fresh source material. *Mol Phylogent Evol* 82:87–94.
Moyle RG, Jones RM, Andersen MJ. 2013. A reconsideration of *Gallicolumba* (Aves: Columbidae) relationships using fresh source material reveals pseudogenes, chimeras, and a novel phylogenetic hypothesis. *Mol Phylogent Evol*. 66:1060–1066.
Mundy NI, Unitt P, Woodruff DS. 1997. Skin from feet of museum specimens as a non-destructive source of DNA for avian genotyping. *Auk* 114:126–129.
Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. *Mol Biol Evol*. 28:2197–2210.
Patel S, Kimball RT, Braun EL. 2013. Error in phylogenetic estimation for bushes in the tree of life. *J Phylogent Evol Biol*. 1:110.
Philippe H. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol*. 21:1740–1752.
Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. *PLoS Biol*. 9:e1000602.

Potapov LR. 2000. New information on the snow partridge *Lerwa lerwa* (Hodgson 1833) and its systematic position. *Bull Br Orn Club*. 120:112–120.
Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526:569–573.
Rannala B, Yang Z. 2008. Phylogenetic inference using whole genomes. *Annu Rev Genomics Hum Genet*. 9:217–231.
Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, Satler JD, Carstens BC. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical Data. *Syst Biol*. 63:322–333.
Roch S, Steel M. 2014. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol*. 100:56–62.
Roch S, Warnow T. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst Biol*. 64:663–676.
Rosenberg NA. 2013. Discordance of species trees with their most likely gene trees: a unifying principle. *Mol Biol Evol*. 30:2709–2713.
Rosenfeld JA, Payne A, DeSalle R. 2012. Random roots and lineage sorting. *Mol Phylogent Evol*. 64:12–20.
Roure B, Baurain D, Philippe H. 2012. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. *Mol Biol Evol*. 30:197–214.
Rubin BER, Ree RH, Moreau CS. 2012. Inferring phylogenies from RAD sequence data. *PLoS One* 7:e33394.
Salichos L, Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
Sanderson MJ, McMahon MM, Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. *BMC Evol Biol*. 10:155.
Sefc KM, Payne RB, Sorenson MD. 2007. Single base errors in PCR products from avian museum specimens and their effect on estimates of historical genetic diversity. *Conserv Genet*. 8:879–884.
Simmons MP. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. *Mol Phylogent Evol*. 80:267–280.
Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Syst Biol*. 63:83–95.
Smith JV, Braun EL, Kimball RT. 2013. Ratite nonmonophyly: independent evidence from 40 novel loci. *Syst Biol*. 62:35–49.
Snir S, Rao S. 2010. Quartets MaxCut: a divide and conquer quartets algorithm. *IEEE/ACM Trans Comput Biol Bioinform*. 7:704–718.
Snir S, Rao S. 2012. Quartet MaxCut: a fast algorithm for amalgamating quartet trees. *Mol Phylogent Evol*. 62:1–8.
Springer MS, Gatesy J. 2016. The gene tree delusion. *Mol Phylogent Evol*. 94:1–33.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML web servers. *Syst Biol*. 57:758–771.
Streicher JW, Schulte JA, Wiens JJ. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. *Syst Biol*. 65:128–145.
Sun K, Meiklejohn KA, Faircloth BC, Glenn TC, Braun EL, Kimball RT. 2014. The evolution of peafowl and other taxa with ocelli (eyespots): a phylogenomic approach. *Proc Biol Soc*. 281:20140823.
Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony, version 4.0b10. Sunderland (MA): Sinauer Associates, Inc.
Thomson RC, Shaffer HB. 2010. Sparse supermatrices for phylogenetic inference: taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Syst Biol*. 59:42–58.
Tonini J, Moore A, Stern DL, Shcheglovitova M, Orti G. 2015. Concatenation and species tree methods exhibit statistically indistinguishable accuracy under a range of simulated conditions. *PLoS Curr* 7.
Townsend JP. 2007. Profiling phylogenetic informativeness. *Syst.Biol*. 56:222–231.
Vachaspati P, Warnow T. 2015. ASTRID: Accurate Species TRees from Internode Distances. *BMC Genomics* 16:S3.
Wagner CE, Keller I, Wittwer S, Selz OM, Mwaiko S, Greuter L, Sivasundar A, Seehausen O. 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. *Mol Ecol*. 22:787–798.
Wang N, Kimball RT, Braun EL, Liang B, Zhang Z. 2013. Assessing phylogenetic relationships among Galliformes: a multigene phylogeny with expanded taxon sampling in Phasianidae. *PLoS One* 8:e64312.
Warnow T. 2015. Concatenation analyses in the presence of incomplete lineage sorting. *PLoS Curr*. 22:7.
Whitfield JB, Lockhart PJ. 2007. Deciphering ancient rapid radiations. *Trends Ecol Evol*. 22:258–265.
Wiens JJ, Morrill MC. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. *Syst Biol*. 60:719–731.
Wiens JJ, Tiu J. 2012. Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling. *PLoS One* 7:e42925.
Zuccon D, Ericson PGP. 2010. The *Monticola* rock-thrushes: phylogeny and biogeography revisited. *Mol Phylogent Evol*. 55:901–910.